

FBSNG - Batch System for Farm Architecture

J. Fromm, K. Genser, T. Levshina, I. Mandrichenko
(Fermi National Accelerator Laboratory, U.S.A.)

Abstract

FBSNG [1] is a redesigned version of Farm Batch System (FBS [2]), which was developed as a batch process management system for off-line Run II data processing at FNAL. FBSNG is designed for UNIX computer farms and is capable of managing up to 1000 nodes in a single farm. FBSNG allows users to start arrays of parallel processes on one or more farm computers. It uses a simplified *abstract resource counting* method for load balancing between computers. The resource counting approach allows FBSNG to be a simple and flexible tool for farm resource management. FBSNG scheduler features include guaranteed and controllable "fair-share" scheduling. FBSNG is easily portable across different flavors of UNIX. The system has been successfully used at Fermilab as well as by off-site collaborators for several years on farms of different sizes and different platforms for off-line data processing, Monte-Carlo data generation and other tasks.

Keywords: batch system, computing farms, PC farms, resource allocation

1 Introduction

Large farms of inexpensive Intel-based computers running Linux OS have become a common tool for off-line data processing at Fermilab as well as other HEP organizations. Estimated size of PC farm used by a Run II experiment is more than 200 dual-CPU 1 GHz Pentium-based computers with about 500 processes concurrently running on the farm, 2-3 processes per computer. FBSNG was developed at Fermilab as a part of farm production management infrastructure toolkit.

Although collider experiments are biggest users of PC farms today, other smaller groups are adopting farm style of computing. FBSNG is designed to manage resource allocation and distribution among different groups of users sharing the same farm as well as among concurrent projects within the same group.

2 Features

2.1 Load Management

FBSNG uses a simplified concept of resource management. Instead of measuring utilization of such resources as CPU, disk space, network load on farm nodes (*load measuring*), FBSNG uses a resource counting algorithm. Farm administrator describes the farm in terms of available *abstract resources*. They are called abstract because the batch system knows only their stated capacity, but nothing else about their nature. A farm user specifies how much resources the batch job will need. FBSNG simply compares job resource requirements to the amount of resources available and decides when and where to allocate requested resources and start the job. This simplified algorithm has proven to be suitable for farm architecture, provides more flexibility than load measuring and makes the batch system simpler and more robust.

2.2 Job Structure

A unit of FBSNG operation is a *batch job*. FBSNG job consists of one or more *job sections*. The user can synchronize sections of a job by defining *dependencies* between them. Each section is an array of identical processes started simultaneously on farm nodes.

2.3 Scheduler

The atomic unit of scheduling is a job section, which is an array of batch processes. The FBSNG scheduler uses an algorithm based on the idea of *dynamic priorities* to provide the following features:

- **fair share scheduling** - the administrator can define what portion of farm resources should be allocated to each group or project
- **guaranteed scheduling** - no matter how big a job is compared to other pending jobs in terms of requested resources, the big job is guaranteed to start within finite amount of time
- **load leveling** - batch processes are evenly placed on farm nodes to equally distribute load over the farm

3 Scalability

FBSNG is designed to manage farms composed of up to 1000 nodes with more than 2000 batch processes running at the same time and process start rate of more than 2000 per hour. If necessary, most of these limits can be increased.

4 Other Features

FBSNG provides a wide range of job control and resource monitoring functions through command line interface, GUI, web interface and Python API. It supports Kerberos5 authentication and creates Kerberos credentials for batch processes. FBSNG is easily portable across various UNIX flavors. Currently it is supported on IRIX, Linux, SunOS and OSF1 platforms.

5 Current Status

FBSNG together with other farm infrastructure tools such as FIPC, Farm data transfer utility (FCP) and Disk Farm [3] has been successfully used on a number of farms at FNAL as well as by off-site collaborators. Known FBSNG-managed farms include CDF [4] and D0 [5] off-line data processing farms, two "fixed target" farms at FNAL, US CMS Tier 1 site [6], D0 farm at NIKHEF [7] and some others.

References

- [1] <http://www-isd.fnal.gov/fbsng/>
- [2] M.Breitung *et al.*, Farm Batch System (FBS) and Fermi Inter-Process Communication Toolkit (FIPC), Preprint FERMILAB-CONF-00/083, CHEP 2000 proceedings
- [3] <http://www-isd.fnal.gov/fcs/>
- [4] S. Wolbers *et al.*, The CDF Run 2 Offline Computer Farms, CHEP 2001 proceedings
- [5] H. Schellman *et al.*, The D0 Offline Farms at Fermilab, CHEP 2001 proceedings
- [6] Y.Wu *et al.*, Simulating the Farm Production System Using the MONARC Simulation Tool, CHEP2001 proceedings
- [7] <http://www.nikhef.nl/~bosk/farm/farm.htm>